University of Illinois
Fall 2016

Department of Economics
Roger Koenker

Economics 536
**Lecture 21**

# Counts, Tobit, Sample Selection, and Truncation

The simplest of this general class of models is Tobin's (1958) model for durable demand

$$y_i^* = x_i^\top \beta + u_i \qquad u_i \sim iid\ F$$

$$y_i = \max\{y_i^*, 0\}$$

That is, we have a propensity, a latent variable, which describes demand for something – when $y_i^* > 0$ we act on it otherwise we do nothing. This model is the simplest form of the *Censored regression* model. The first question we should address is Why not estimate by OLS? First, we must clarify: "OLS on what?" Let's consider OLS on just the $y_i > 0$ observations. Recall that OLS tries to estimate the conditional mean function for $y$ so let's try to compute this in our case:

$$y_i^* = x_i\beta + u_i$$

so

$$E(y_i|y_i^* > 0) = x_i^\top \beta + E(u_i|y_i^* > 0) = x_i^\top \beta + E(u_i > -x_i^\top \beta)$$

by the Appendix A

$$= x_i\beta + \frac{\sigma\phi(x_i^\top \beta/\sigma)}{\Phi(x_i^\top \beta/\sigma)}$$

when $u_i \sim iid\ \mathcal{N}(0, \sigma^2)$. Thus

$$E\hat{\beta} = (X^\top X)^{-1}X^\top Ey = \beta + \sigma(X^\top X)^{-1}X^\top \lambda$$

where $\lambda = (\phi_i/\Phi_i)$. Note that all the mass corresponding to $y^* < 0$ piles up at $y = 0$. So we get a nonlinear conditional expectation function.

## 1. The Heckman 2-step Estimator

This suggests that if we could somehow estimate $\beta/\sigma = \gamma$ we might be able to correct for the bias introduced by omitting the zero observations. How to estimate $\gamma$? The tobit model as expressed above is just the probit model we have already considered except that in the previous case $\sigma \equiv 1$, but note here we can divide through by $\sigma$ in the first equation without changing anything. Then it is clear that we are estimating $\gamma = \beta/\sigma$ by the usual probit estimator. So Heckman(1979) proposes:

(1.) Estimate binary choice model by probit.

(2.) Construct $\hat{\lambda}_i = \phi(x_i^\top \gamma)/\Phi(x_i^\top \hat{\gamma})$.

(3.) Reestimate original model using only $y_i > 0$ observations but including $\hat{\lambda}_i$ as additional explanatory variable the coefficient estimated on $\lambda$ is $\sigma$.

This approach is helpful because it clarifies what is going wrong in OLS estimation and how to correct it, but it is problematic in several other respects. In particular, it is difficult to construct s.e.'s

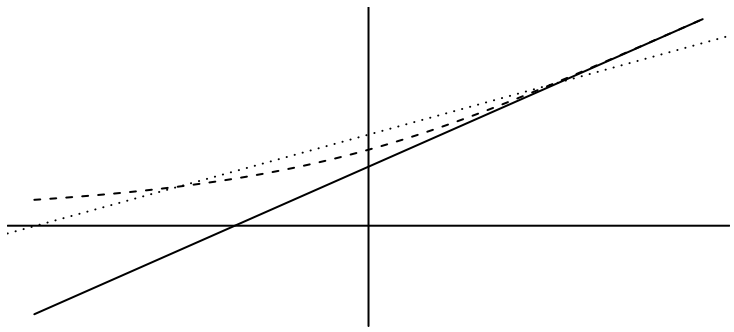

FIGURE 1. Bias of OLS estimator in the Censored Regression Model: The figure illustrates the conditional expectation of the latent variable $y_i^*$ given $x$ as the solid straight line in the figure. The conditional expectation of the observed response $y_i$ is given by the curved dotted line. And the least squares linear approximation of the conditional expectation of the observed response is given by the dashed line. Note that in this model the conditional median function of $y_i$ given x is the piecewise linear function $\max\{a + bx, 0\}$, where $E(y_i^*|x) = a + bx$.

for the estimates since the effect of the preliminary estimate of $\gamma$ is non-negligible. It is also instructive to consider the mle in this problem. The likelihood is straightforward to write down:

$$\mathcal{L}(\beta, \sigma) = \prod_{i:y_i=0} F\left(-\frac{x_i^\top \beta}{\sigma}\right) \prod_{i:y_i>0} \sigma^{-1} f((y_i - x_i^\top \beta)/\sigma)$$

for $F = \Phi$ we have

$$= \prod_{i:y_i=0} (1 - \Phi\left(\frac{x_i^\top \beta}{\sigma}\right)) \prod_{i:y_i>0} \sigma^{-1} \phi((y_i - x_i^\top \beta)/\sigma)$$

It is useful to contrast this censored regression estimator with the truncated regression estimator with likelihood,

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^{n} (\Phi(x_i^\top \beta/\sigma))^{-1} \phi((y_i - x_i^\top \beta)/\sigma)$$

## 2. Powell's estimator

A critical failing of the Gaussian mle is that it can perform poorly in non-Gaussian and/or heteroscedastic circumstances. If we go back to our picture we can see that the primary source of the difficulty we have been discussing is due to the wish to estimate conditional *expectations*. If, instead, we tried to estimate the condition median then we have

$$(*) \qquad \text{med}(y_i|x_i) = \max\{x_i^\top \beta, 0\}$$

so we can following Powell (1984) consistently estimate $\beta$ by solving

$$\min \sum |y_i - \max\{x_i^\top \beta, 0\}|$$

This works for any $F$ as long as (*) holds, *even if there is heteroscedasticity.* This can be easily extended (Powell (1985) to quantile regression in general. An interesting question is what quantiles offer optimal efficiency in estimating $\beta$. The computational difficulty of this approach is substantially greater than conventional quantile regression due to the fact that the objective function is no longer convex. This problem has been addressed by several authors, notably Fitzenberger (1997). Chernozhukov and Hong (2000) have suggested an alternative simpler approach that employs a multi-step procedure that has the advantage that it only requires solution to linear-in-parameters quantile regression problems. Portnoy (2003) has introduced a new method for fitting randomly censored quantile regression models.

## Models for Count Data

Often we have data that is integer valued and nonnegative: number of doctors visits per year, or number of patents awarded per year, etc. A natural first choice for such data is the Poisson model.

$$P(Y_i = k) = f(k) = e^{-\lambda}\lambda^k/k! \qquad k = 0, 1, 2, \ldots$$

Given covariates, we would like to formulate a regression type model, so we may take

$$\lambda = e^{x_i'\beta}$$

The log likelihood for the parameter $\beta$ given a sample $(x_i, y_i : i = 1, \ldots, n)$. is,

$$\ell(\beta) = \sum y_i x_i^\top \beta - exp(x_i^\top \beta) - \log(y_i!)$$

and maximizing leads us to the "normal" equations,

$$\sum (y_i - exp(x_i^\top \beta))x_i = 0$$

Again, we have a system of nonlinear equations to solve in $\beta$, but the equations are rather benign, and can be again solved by iteratively reweighted least squares methods as with probit and logit.

In a now classic (i.e. forgotten) study of meta-econometrics Koenker (1988) estimated a model of this type purporting to explain the number of parameters $k_i$, as a function of the sample size $n_i$ of the study, estimated in published studies of wages in the labor economics literature. Two of the estimated models were:

$$(1) \qquad \begin{array}{rcccc} \log \lambda_i & = & 1.34 & + & .235 \quad \log n_i \\ & & (.14) & & (.017) \end{array}$$

$$(2) \qquad \begin{array}{rcccc} \log \lambda_i & = & -0.78 & + & 1.947 \quad \log(\log n_i) \\ & & (.31) & & (.148) \end{array}$$

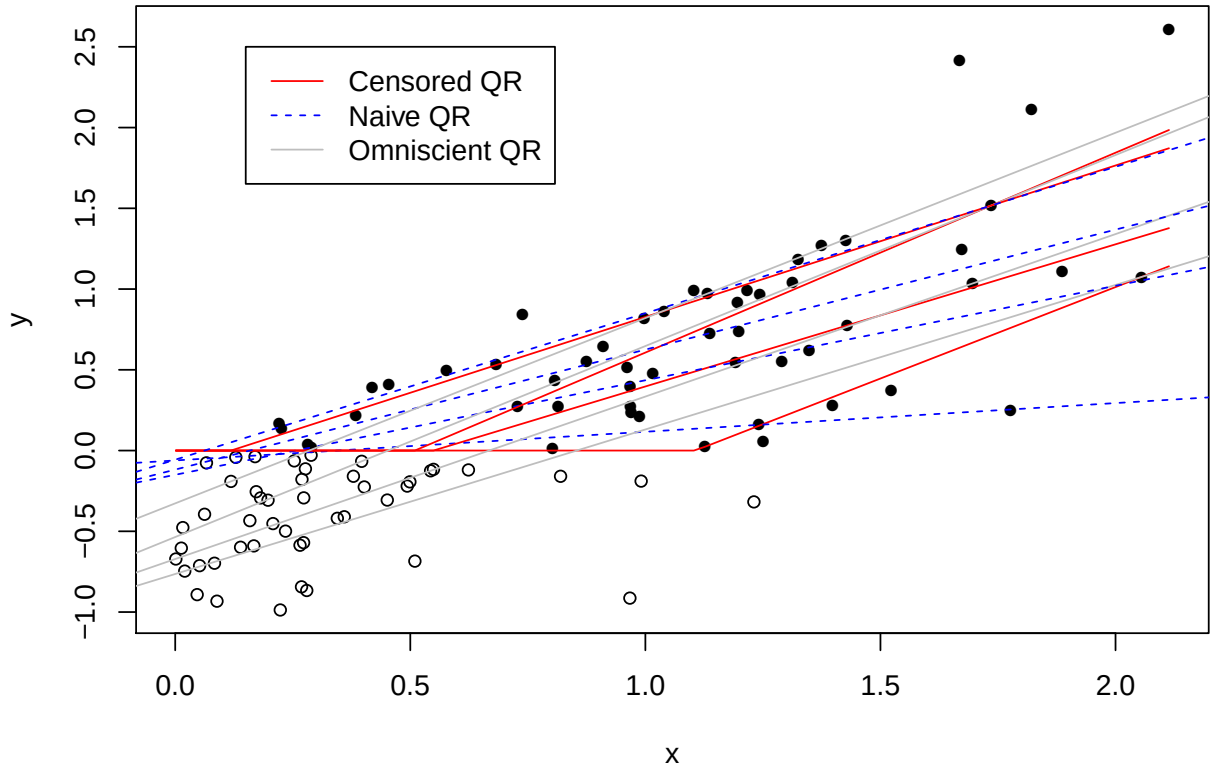


FIGURE 2. Powell estimator: The figure illustrates three families of estimated quintile lines to the points appearing in the plot. In accordance with the classical tobit model, the points with $y$ less than zero are censored so they are recorded as zero. The dashed lines illustrate the naive QR fit that ignores the fact of the censoring, the grey lines represent the omniscient version of the fitting obtained if we "know" the latent values of the censored observations, and the solid black lines are the estimated Powell quintiles. Note that the Powell estimates are quite close to omniscient estimates as we might hope, except when the estimated line bends to reflect the censoring at zero. This plot is essentially what is obtained by running `example(rq.fit.cen)` in the quantreg package of R. The algorithm of Fitzenberger is employed for the Powell fitting.

Mean sample size for the sample of wage equations was about $\bar{n} = 1000$. If one were mainly interested in the effect,

$$\pi = \frac{\partial \log Ek}{\partial \log n}.$$

which we might call the parsity, or elasticity of parsimony for lack of better terms, how different are the two models? Recall that expectation of a Poisson random variable with rate $\lambda$ is simply $\lambda$. Contrast the implications of the two models with respect to the behavior of $E(k_i|n_i)$ as $n_i \to \infty$.

The Poisson model has the feature that it assumes that the variance of a Poisson variable is the same as its mean. This may be sensible in some applications, but uncritical acceptance of this hypothesis sometimes leads to very unreliable inference. Classical MLE theory suggests that the variance covariance of $\hat{\beta}_n$ solving the normal equations is

$$V_{MLE} = (\sum \hat{\lambda}_i x_i x_i^\top)^{-1}$$

adhering to this assumption, but it is safer to admit that the the variance may depend on the mean in some more general way. When the Poisson assumption fails a general form for the covaraince matrix of $\hat{\beta}_n$ is,

$$V_{MLE} = (\sum \hat{\lambda}_i x_i x_i^\top)^{-1} (\sum \hat{\nu}_i x_i x_i^\top)(\sum \hat{\lambda}_i x_i x_i^\top)^{-1}$$

where $\hat{\nu}_i$ is an estimate of $V(y_i|x_i)$. There is a good discussion of various strategies for estimating this conditional variance in the monograph of Cameron and Trivedi (1998).

## Simple Heckman Sample Selection Model

Now, we will extend the tobit model to a somewhat more general setup which is usually associated with a labor supply model of Gronau. Gonsider two latent variable equations,

$$\begin{aligned} y_1^* &= x_1^\top \beta_1 + u_1 \\ y_2^* &= x_2^\top \beta_2 + u_2 \end{aligned}$$

and assume that we observe

$$y_1 = \begin{cases} 1 & y_1^* > 0 \\ \text{if} & \\ 0 & y_1^* \leq 0 \end{cases} \qquad y_2 = \begin{cases} y_2^* & y_1 = 1 \\ \text{if} & \\ 0 & y_1 = 0 \end{cases}$$

where in the labor supply model $y_1$ may be interpreted as the decision to enter the labor force and $y_2$ is the number of hours worked. Then,

$$E(y_2|x_2, y_1 = 1) = x_2^\top \beta_2 + E(u_2|u_1 > -x_1^\top \beta_1)$$

but by Appendix B $u_2|u_1 \sim \mathcal{N}(\frac{\sigma_{12}}{\sigma_1^2} u_1, \sigma_2^2 - \sigma_{12}^2 \sigma_1^{-2})$ so

$$E(y_2|x_2, y_1 = 1) = x_2^\top \beta_2 + E(\frac{\sigma_{12}}{\sigma_1^2} u_1 | u_1 > -x_1' \beta_1)^\top$$

Recall from Tobit case

$$E(u_1|u_1 > -x_1^\top \beta_1) = \frac{\sigma_1 \phi(x_1^\top \beta_1/\sigma_1)}{\Phi(x_1^\top \beta_1/\sigma_1)} = \sigma_1 \lambda$$

so

$$E(y_2|x_2, y_2 = 1) = x_2^\top \beta_2 + \frac{\sigma_{12}}{\sigma_1} \lambda(x_1^\top \beta_1/\sigma_1)$$

which may now be estimated by Heckman 2-step as follows.

(1.) Probit of $y_1$ on $x_1$ to get $\hat{\gamma}$ if $\beta_1/\sigma_1$.

(2.) Construct $\hat{\lambda}$ and regress $y_2$ on $[X_2 \vdots \hat{\lambda}]$.

(3.) Test for Sample Selection bias using $\sigma_{12}/\sigma_1$ estimate. Or, this could be estimated via mle methods.

Increasingly, researchers have grown dissatisfied with the Heckman latent variable model recognizing that under misspecification of either the normality assumption or due to various forms of heterogeneity large biases may ensue. Manski (1989) offers a radical reappraisal of the problem. He begins with the observation that we can write,

$$P(y|x) = P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x) \tag{1}$$

when $z$ denotes the binary selection variable. We would like to know $P(y|x)$, but since $P(y|x, z = 0)$ is unobserved – we don't know for example what wages are like for the unemployed – there is a fundamental identification problem. This can be addressed in various parametric ways. The simplest of these is to assume selection away. It turns out to be particularly difficult to identify mean response given general assumptions for (1). In contrast quantiles of $y$ are somewhat more tractable. Let

$$\hat{Q}_y(\tau|x) = \inf \{\xi | P(y \geq \xi|x) \geq \tau\}$$

and define

$$\begin{aligned} \underline{Q}_y(\tau|x) &= \begin{cases} Q_y(1 - (1-\tau)/P(z=1|x)|x, z=1) & \text{if} \quad P(z=1|x) >= 1-\tau \\ -\infty \quad \text{otherwise} \end{cases} \\ \bar{Q}_y(\tau|x) &= \begin{cases} Q_y(\tau/P(\tau/P(z=1|x)|x, z=1) & \text{if} \quad P(z=1|x) >= \tau \\ \infty \quad \text{otherwise} \end{cases} \end{aligned}$$

Then, Manski shows that

$$\underline{Q}_y = (\tau|x) \leq Q_y(\tau|x) \leq \bar{Q}_y(\tau|x)$$

The upper and lower bounds are increasing in $\tau$. As along as $P(z = 1|x) < \min\{tau, 1-\tau\}$ the bounds are informative about the $\tau$th quantile. The implication of this, of course, is that we can't bound quantiles in the tails and therefore we can't bound the mean effect.

A truly semi-parametric approach to sample selection has been an elusive quest in econometrics for quite some time. Recent unpublished work by Arellano and Bonhomme offers hope of resolving this in a positive way. Stay tuned.

## References


Fitzenberger, B. (1997) A Guide to Censored Quantile Regressions, C.R. Rao and G.S. Maddala (eds.),*Handbook of Statistics*, North-Holland: New York,

Cameron, C. and P. Trivedi (1998) Regression Analysis for Count Data, Cambridge U. Press.

Chernozhukov, V. and H. Hong, (2000) 3-step Censored Quantile Regression, JASA, 97,872–897.

Manski, C.F. (1989) Anatomy of the selection problem, *J. of Human Resources*, **21**, 343-360.

Manski, C.F. (1993) The selection problem in econometrics and Statistics, *Handbook of Statistics*, **11**, 73-83.

Tobin, J. (1958) Estimation of Relationships for limited dependent variables, *Econometrica*, **26**, 24-36.

Heckman, J. (1979) Sample Selection as a Specification Error, *Econometrica*, **47**, 153-62.

Portnoy, S. (2003). Censored Regression Quantiles, *J. Am. Stat. Assoc*, 98, 1001-1012.

Powell, J.L. (1984). Least absolute deviations estimation for the censored regression model, *J. Econometrics, 25*, 303-325.

Powell, J.L. (1986). Censored regression quantiles, *J. Econometrics, 32*, 143-155.

## APPENDIX A:
## Some Notes on Conditional Expectations for the Tobit Model

If $Z$ has $dfF$ with density $f$, then the conditional density of $Z$ given $Z > c$ is

$$f_c(z) = f(z)/(1 - F(c))$$

Note

$$\int_c^\infty f_c(z)dz = (1 - F(c))^{-1} \int_c^\infty f(z)dz = 1$$

as expected. The condition expectation of $Z$ given $Z > c$ is

$$E(Z|Z > c) \quad = \quad \int z f_c(z)dz = (1 - F(c))^{-1} \int_c^\infty z f(z)dz.$$

For $F$ standard Gaussian we have $zf(z) = z\phi(z) = -\phi'(z)$ so,

$$\begin{aligned} E(Z|Z > c) &= (1 - \Phi(c))^{-1}(-\int_c^\infty \phi'(z)dz) \\ &= \phi(c)/(1 - \Phi(c)). \end{aligned}$$

Finally, consider $Y = \sigma Z$ so $Y \sim \mathcal{N}(0, \sigma^2)$.

$$\begin{aligned} E(Y|Y > c) &= E(\sigma Z|\sigma Z > c) \\ &= \sigma E(Z|Z > c/\sigma) \\ &= \sigma\phi(c/\sigma)/(1 - \Phi(c/\sigma)). \end{aligned}$$

## APPENDIX B
## Conditional Normality

*Theorem:* Let $Y$ be $p$-variate normal $\mathcal{N}(\mu, \Omega)$ with sub-vectors $Y_1$ and $Y_2$ having $EY_i = \mu_i$, and $\text{Cov}(Y_i, Y_j) = \Omega_{ij}$. Assume $\Omega_{11}$ and $\Omega_{22}$ are nonsingular. Then the conditional distribution of $Y_2$ given $Y_1$ is $\mathcal{N}(\mu_2 + \Omega_{21}\Omega_{11}^{-1}(Y_1 - \mu_1), \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})$.

*Proof:* (This is a simplified version of Rao, Linear Stat Inference, 1973, p. 523. Rao relaxes the nonsingularity condition.) Consider

$$\text{Cov}[Y_2 - \mu_2 - \Omega_{21}\Omega_{11}^{-1}(Y_1 - \mu_1), Y_1 - \mu_1] = \Omega_{21} - \Omega_{21}\Omega_{11}^{-1}\Omega_{11} = 0 \qquad (*)$$

Similarly, let $U = Y_2 - \mu_2 - \Omega_{21}\Omega_{11}^{-1}(Y_1 - \mu_1)$ clearly $EU = 0$ and

$$V(U) = V[Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1]$$

$$= \Omega_{22} + \Omega_{21}\Omega_{11}^{-1}\Omega_{12} - \text{Cov}(Y_2, \Omega_{21}\Omega_{11}^{-1}Y_1) - \text{Cov}(\Omega_{21}\Omega_{11}^{-1}Y_1, Y_2)$$

$$= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}$$

Since $U$ is a linear function of normal *r.v.*'s it is normal, and therefore,

$$U \sim \mathcal{N}(0, \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}) \qquad (+)$$

Further, $(*)$ establishes that $U$ and $Y_1 - \mu_1$ are independent, hence $(+)$ may be interpreted as the conditional distribution of $U$ given $Y_1$, which is equivalent to what we wished to prove.